# How Much Did It Rain?

**Zhizhong Li**
Oct 5, 2016
Information Engineering, CUHK
1155070507, lz015@ie.cuhk.edu.hk

## 1 Introduction

This report use sketching methods to deal with the least square problem which minimizes

$$f(x) = \|Ax - b\|_2, \tag{1}$$

where $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $x \in \mathbb{R}^d$. The four sketching methods are *Gaussian*, *PHD*, *Count Sketch*, and *Leverage Score*.

### 1.1 Dataset

We use the *how much did it rain ii* [1, 2] dataset from *kaggle.com* for the analysis of different sketching methods. The original dataset contains 13,765,201 training samples with 23 features and 1 prediction value. The following description was extracted from web page [1].

**Id** A unique number for the set of observations over an hour at a gauge.

**Minutes_past** For each set of radar observations, the minutes past the top of the hour that the radar observations were carried out. Radar observations are snapshots at that point in time.

**Radardist_km** Distance of gauge from the radar whose observations are being reported.

**Ref** Radar reflectivity in km.

**Ref_5x5_10th, Ref_5x5_50th, Ref_5x5_90th** 10th, 50-th, 90-th percentile of reflectivity values in 5x5 neighborhood around the gauge.

**RefComposite, RefComposite_5x5_10th, RefComposite_5x5_50th, RefComposite_5x5_90th** Maximum reflectivity in the vertical column above gauge. In dBZ.

**RhoHV, RhoHV_5x5_10th, RhoHV_5x5_50th, RhoHV_5x5_90th** Correlation coefficient.

**Zdr, Zdr_5x5_10th, Zdr_5x5_50th, Zdr_5x5_90th** Differential reflectivity in dB.

**Kdp, Kdp_5x5_10th, Kdp_5x5_50th, Kdp_5x5_90th** Specific differential phase (deg/km).

**Expected** Actual gauge observation in mm at the end of the hour.

Since it contains null values and outliers, we will do a preprocess and only use a subset of samples. All the experiments were conducted on a machine that has an 16-core Intel Xeon E5-2637 v2 CPU at 3.5GHz, and 256G memory. Codes were implemented using the Julia language and are available at *https://github.com/innerlee/random.computation.report1*.

### 1.2 Preprocess and Baselines

Since some observations are not complete, we first filter out 2,769,088 samples that do not contain missing values. The first two features *id* and *minutes_past* are used just for identification rows, thus we omit them and get data matrices

$$A_0 \in \mathbb{R}^{2769088\times 21},\ b_0 \in \mathbb{R}^{2769088}. \tag{2}$$

The distribution of these inputs was shown in Figure 1.

This is the *Randomness and Computation* HW1 report.

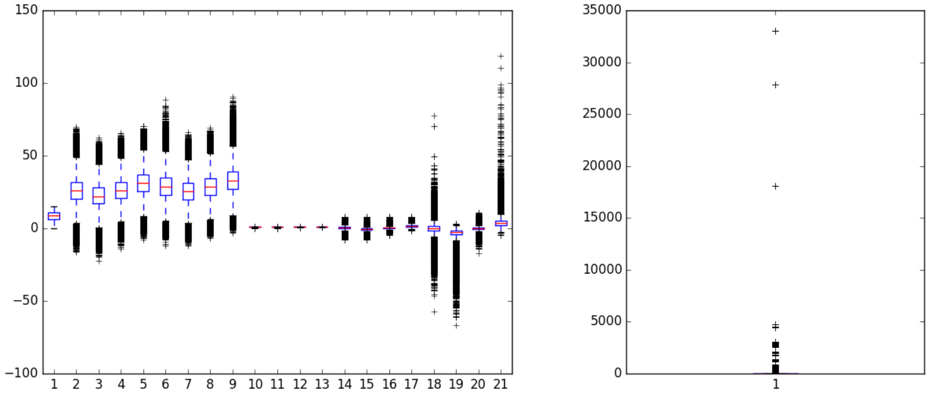

Figure 1: Left, box plot for columns of $A_0$. Right, that for $b_0$. We can see $b_0$ contains outliers.

Firstly, direct compute the least square problem (1) by equation

$$x_0^* = (A_0^T A_0)^{-1} A_0^T b_0. \tag{3}$$

We get

$$x_0^* = (0.758, 0.046, 0.218, \ldots, -0.035, -0.101)$$

with average loss $f(x_0^*)/\sqrt{n} = 156.1$ in 0.19 seconds. Compared to mean$(b_0) = 12.2$, the mean value of $b_0$, the average loss seems too large due to outliers. The 99-th percentile of prediction values in $b_0$ is 144.0. We then use samples that have prediction values less than 144.0 and get new data

$$A \in \mathbb{R}^{2741220\times 21},\ b \in \mathbb{R}^{2741220}. \tag{4}$$

Directly compute $x^*$ using Equation (3) again, we can get $x^* = (0.189, -0.012, -0.008, 0.022, 0.13, -0.029, -0.063, 0.081, 0.076, 0.061, 0.222, -2.724, -1.046, 0.015, 0.082, 0.092, -0.138, 0.004, 0.027, 0.042, -0.04)$ with loss $f(x^*) = 15090.7$ in 0.19 seconds. This makes more sense because the average loss $f(x_0^*)/\sqrt{n} = 9.11$ is comparable to stats like mean$(b) = 4.24$, std$(b) = 9.30$, minimum$(b) = 0.01$, and maximum$(b) = 142.2$. And this serves as the baseline for our later discussion.

## 2 Sketching

Let $n = 2741220$ and $d = 21$, which are taken from dimensions of data in Equation (4). Set $\varepsilon = 0.1$ and $\delta = 0.9$ in the following. We gather all the testing results in Table 1.

Table 1: The performances of different Sketching techniques. other than the min loss, all are average values of repeats. *prep* and *app* are time spent on sketching and on solving the shrinked sized least square problem, respectively. ref-$k$ is the reference value for $k$.

| algorithm | repeat | ref-$k$ | $k$ | prep (s) | app (ms) | min loss | median loss | max loss | std loss | mean loss |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 10 | - | - | 0 | 190 | 15090.7 | 15090.7 | 15090.7 | 0 | 15090.7 |
| Gaussian | 100 | 2432 | 10 | 0.76 | 0.3 | 20479.8 | 76514.7 | 1268090 | 177191.0 | 126311 |
| Gaussian | 100 | 2432 | 100 | 7.44 | 9.7 | 16080.5 | 16983.8 | 18631.3 | 583.3 | 17059.5 |
| PHD | 100 | 37234 | 100 | 3.76 | 0.4 | 15927.6 | 16977.5 | 19309.7 | 635.3 | 16999.5 |
| PHD | 100 | 37234 | 1000 | 3.72 | 1.1 | 15159.7 | 15252.5 | 15364.3 | 45.4 | 15248.1 |
| Count | 100 | 340198 | 1000 | 1.17 | 0.4 | 15157.6 | 15240.8 | 15397.1 | 43.0 | 15246.7 |
| Count | 100 | 340198 | 10000 | 1.20 | 1.1 | 15096.8 | 15105.3 | 15122.3 | 5.3 | 15106.5 |
| Leverage | 100 | 1826573 | 10000 | 2.81 | 1.3 | 15093.7 | 15096.8 | 15105.4 | 2.2 | 15097.3 |
| Leverage | 100 | 1826573 | 100000 | 4.89 | 9.4 | 15091.0 | 15091.3 | 15092.1 | 0.2 | 15091.4 |

### 2.1 Gaussian

By Theorem 6 in the text book article [3],

$$k = \Theta((d + \log(1/\delta))\varepsilon^{-2}), \tag{5}$$

$k$ is a multiple of $(21 + log(10)) \times 100 \approx 2432$. This is too large for real computation, so we set $k = 10$ and $k = 100$. Results are shown in Table 1. The minimum point when $k = 100$ is $x^* = (0.292, 0.162, 0.83, -1.478, 0.693, 0.034, -0.851, 1.482, -0.614, 1.286, -29.447, -12.258, 37.858, 0.534, 5.043, -5.779, -0.274, -0.157, -0.212, 0.753, -0.234)$ with loss $f(x^*) = 16080.5$.

### 2.2 PHD

$P$ is used for pick $k$ rows from matrix on the right hand side. $H$ is Hardamard matrix and $D$ a diagonal matrix. Since size of $H$ is required to be a power of 2, we pad the the $n$ dimensional vectors to the smallest power of 2 that is larger than $n$. We use *Hadamard.jl* package in Julia for the computation of the Hadamard transform. By Theorem 7, the selected row number

$$k = \Omega\big((\log(d))(\sqrt{d} + \sqrt{\log(n)})^2\varepsilon^{-2}\big). \tag{6}$$

A reference number for $k$ is $\log(21) \times (\sqrt{21} + \sqrt{\log(2741220)})^2 \times 100 \approx 37234$. This is too large for computation, thus we set $k = 1000$ and $k = 1000$. Results are shown in Table 1. The minimum point when $k = 1000$ is $x^* = (0.221, 0.002, 0.021, 0.193, -0.034, -0.048, 0.027, -0.168, 0.24, -1.341, -4.14, -9.779, 10.611, -0.129, -0.359, 1.604, -0.894, -0.035, -0.121, 0.28, -0.065)$ with loss $f(x^*) = 15159.7$.

### 2.3 Count Sketch

By Theorem 8, the $k$ (we use $k$ here for consistency, $r$ is used in the article) is,

$$k = \mathcal{O}\big(d^2 \text{poly}(\log(d/\varepsilon))\varepsilon^{-2}\big). \tag{7}$$

A reference number for $k$ is $21^2 \times \log(21/0.1) \times 100 \approx 340198$. This is too large for computation, thus we set $k = 1000$ and $k = 10000$. Results are shown in Table 1. The minimum point when $k = 10000$ is $x^* = (0.179, -0.007, 0.004, 0.036, 0.09, 0.007, -0.09, 0.083, 0.081, -0.871, -1.128, 4.666, -6.164, 0.127, 0.214, -0.183, -0.245, 0.011, 0.004, 0.02, 0.002)$ with loss $f(x^*) = 15096.8$.

### 2.4 Leverage Score

We use the procedure described by Definition 16. For simplicity, we did not implement the fancier one as in Theorem 19. In this case, $q$ is selected as $p$ and $\beta = 1$. By Theorem 17, the $k$ (we use $k$ here for consistency, $s$ is used in the article, and the $k$ of the article correspond to $d$ here) is,

$$k > 144 d \ln(2d/\delta)\beta^{-1}\varepsilon^{-2}. \tag{8}$$

A reference number for $k$ is $144 \times 21 \times \ln(2 \times 21 \times 10) \times 100 \approx 1826573$. This is too large for computation, thus we set $k = 10000$ and $k = 100000$. Results are shown in Table 1. The minimum point when $k = 10000$ is $x^* = (0.193, -0.021, -0.004, 0.031, 0.128, -0.024, -0.078, 0.086, 0.078, -0.099, -0.079, -2.097, -1.301, 0.015, 0.066, 0.091, -0.136, 0.001, 0.028, 0.046, -0.034)$ with loss $f(x^*) = 15091.0$. This is the best result in our experiment. Especially, notice that both $x^*$ and $f(x^*)$ are close to the ground truth.

## 3 Discussion

From the experiment results, we can see that

- The preparation time for the sketching is large in experiments. Efforts should be made to reduce this time. The second step of solving the smaller sized problem take negligible amount of time if $k$ is small.
- The performance of the four sketching are Leverage Score, Count Sketch, PHD, and Gaussian in decreasing order.

- The loss can be very near to the ground truth. Especially for the Leverage Score method.
- When $k$ is very large, *e.g.* $100,000$, the std of loss is very small. This means we can get a very good guess stably.

Therefore sketching methods is promising in large scale problems.

## References

[1] Kaggle.com. How much did it rain? ii, oct 2016. URL `https://www.kaggle.com/c/how-much-did-it-rain-ii/`.

[2] V. Lakshmanan, A. Kleeman, J. Boshard, R. Minkowsky, and A. Pasch. The ams-ai 2015-2016 contest: Probabilistic estimate of hourly rainfall from radar. In *13th Conference on Artificial Intelligence*, Phoenix, AZ, 2015. American Meteorological Society.

[3] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.